

Empirical Uncertain Bayes Methods in Area-level Models*

SHONOSUKE SUGASAWA

Risk Analysis Research Center, The Institute of Statistical Mathematics

TATSUYA KUBOKAWA

Faculty of Economics, University of Tokyo

KOTA OGASAWARA

Department of Industrial Engineering and Economics, Tokyo Institute of Technology

Abstract. Random effects model can account for the lack of fitting a regression model and increase precision of estimating area-level means. However, in case that the synthetic mean provides accurate estimates, the prior distribution may inflate an estimation error. Thus it is desirable to consider the uncertain prior distribution, which is expressed as the mixture of a one-point distribution and a proper prior distribution. In this paper, we develop an empirical Bayes approach for estimating area-level means, using the uncertain prior distribution in the context of a natural exponential family, which we call the empirical uncertain Bayes (EUB) method. The regression model considered in this paper includes the Poisson-gamma and the binomial-beta, and the normal-normal (Fay-Herriot) model, which are typically used in small area estimation. We obtain the estimators of hyperparameters based on the marginal likelihood by using a well-known Expectation-Maximization algorithm, and propose the EUB estimators of area means. For risk evaluation of the EUB estimator, we derive a second-order unbiased estimator of a conditional mean squared error by using some techniques of numerical calculation. Through simulation studies and real data applications, we evaluate a performance of the EUB estimator and compare it with the usual empirical Bayes estimator.

Key words: Binomial-beta model, conditional mean squared error, Fay-Herriot model, mixed model, natural exponential family with quadratic variance function, Poisson-gamma model, small area estimation, uncertain random effect.

1 Introduction

In the small area estimation methodology, the model-based empirical Bayes approach has become very popular to produce indirect estimates since a design-based direct estimate of a small area characteristic based on the area sample is usually less reliable because the size of the area sample is small. Importance and usefulness of model-based small area estimation approach may be assessed from an explosive growth of research publications. For a recent comprehensive review of this literature, see Rao and Molina (2015), Jiang and Lahiri (2006), and Pfeiffermann (2013). Random effects models can account for the lack of fit of a regression model and increase precision of estimation for small area mean. The representative model for estimating area means is the Fay-Herriot model (Fay and Herriot, 1979)

*This version : September 12, 2016

described as

$$y_i = \theta_i + \varepsilon_i, \quad \theta_i = \mathbf{x}_i^t \boldsymbol{\beta} + v_i, \quad i = 1, \dots, m,$$

where θ_i corresponds to an area mean, a parameter of interest, ε_i is a sampling error distributed as $\varepsilon_i \sim N(0, D_i)$ for known D_i , and v_i is a random effect distributed as $v_i \sim N(0, A)$ with unknown scalar parameter A .

Recently, however, Datta, Hall and Mandal (2011) showed that the model parameters and the small area means may be estimated with substantially higher accuracy if the random effects can be dispensed with via a suitable test. In response to this work, Datta and Mandal (2015) proposed the Fay-Herriot model with uncertain random effects:

$$y_i = \theta_i + \varepsilon_i, \quad \theta_i = \mathbf{x}_i^t \boldsymbol{\beta} + s_i v_i, \quad i = 1, \dots, m, \quad (1)$$

for $P(s_i = 1) = p = 1 - P(s_i = 0)$. In Datta and Mandal (2015), the term $s_i v_i$ is called “uncertain random effect” since the distribution of $s_i v_i$ is expressed as a mixture of $N(0, A)$ and the one-point distribution on 0. In this setting, the marginal density of y_i is

$$f(y_i) = \frac{1}{\sqrt{2\pi}} \left\{ \frac{p}{\sqrt{A + D_i}} \exp \left(-\frac{(y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2}{2(A + D_i)} \right) + \frac{1-p}{\sqrt{D_i}} \exp \left(-\frac{(y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2}{2D_i} \right) \right\},$$

which is a mixture density of the marginal density of the usual Fay-Herriot model and the simple regression model without random effects. The resulting Bayes estimator (predictor) of θ_i is given by

$$\hat{\theta}_i = \mathbf{x}_i^t \boldsymbol{\beta} + \frac{r_i A}{A + D_i} (y_i - \mathbf{x}_i^t \boldsymbol{\beta}), \quad r_i = p \left\{ p + (1-p) \sqrt{\frac{A + D_i}{D_i}} \exp \left(-\frac{A(y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2}{2D_i(A + D_i)} \right) \right\}^{-1}.$$

It is noted that if $p = 1$, it follows $r_i = 1$ in which $\hat{\theta}_i$ corresponds to the usual linear estimator of θ_i . It is worth noting that the shrinkage coefficient $r_i A / (A + D_i)$ depends on y_i as well as D_i , and r_i increases as the squared residual $(y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2$ increases. Thus if the squared residual is small, which corresponds to the situation that y_i is well explained by \mathbf{x}_i , the shrinkage coefficient gets smaller and $\hat{\theta}_i$ becomes close to the synthetic mean $\mathbf{x}_i^t \boldsymbol{\beta}$. Therefore, the use of the uncertain random effect enables us to take the information about how much \mathbf{x}_i can explain y_i into account for estimating θ_i .

While Datta and Mandal (2015) focused on extending the traditional Fay-Herriot model, we often encounter count data in real applications, and the model based on normal distributions is apparently inappropriate in such a case. Thus, in this paper, we develop a method in a family of distributions called the natural exponential family with quadratic variance functions (NEF-QVF), which includes a normal distribution as well as binomial and Poisson distributions as a special case. This family of distributions was originally investigated in Morris (1982, 1983), and the small area models based on this family were developed by Ghosh and Maiti (2004). To implement the idea of uncertain random effects in this context, we rewrite the uncertain random effect model (1) as the hierarchical form:

$$y_i | \theta_i \sim N(0, D_i), \quad \theta_i | (s_i = 1) \sim N(\mathbf{x}_i^t \boldsymbol{\beta}, A), \quad \theta_i | (s_i = 0) = \mathbf{x}_i^t \boldsymbol{\beta}.$$

This means that one of the two distributions $N(\mathbf{x}_i^t \boldsymbol{\beta}, A)$ and the one point distribution on $\mathbf{x}_i^t \boldsymbol{\beta}$ is randomly selected for the prior distribution of θ_i in each area. In this paper, we naturally extend the idea to the NEF-QVF family. Since there exist the conjugate priors for the natural parameter θ_i , we introduce the uncertain prior for θ_i , a mixture distribution of the conjugate prior and the one-point distribution on the synthetic mean.

For estimating area means under the model, we here develop an empirical Bayes (EB) approach while Datta and Mandal (2015) considered a hierarchical Bayes (HB) approach. In the normal case as

considered in Datta and Mandal (2015), a full Bayesian approach is relatively attractive since all the full conditional distribution of the model parameters as well as the random effects have familiar forms, so that we can efficiently sample from the posterior distribution using a Gibbs sampling. However, in the non-normal case, the posterior distribution of the model parameters are not necessarily in familiar forms, so that we need to rely on an inefficient sampling algorithm such as a Metropolis-Hastings algorithm. Moreover, the HB approach requires checking prior sensitivity and monitoring the convergence of the MCMC algorithm. As suggested in Datta and Mandal (2015), the use of non-informative (improper) enables us to avoid subjective specification of priors, but the posterior propriety is not straightforward under non-normal cases. On the other hand, the empirical Bayes approach can enjoy easily computing point estimates of model parameters and Bayes estimator without requiring prior distributions. Since one of the greatest purposes in small area estimation is point estimation, the EB approach is more attractive in this case.

Owing to the conjugacy of the prior distribution, we can easily establish the Expectation-Maximization (EM) algorithm for maximizing the marginal likelihood function to get the estimates of model parameters. Using the estimator, we derive the the empirical uncertain Bayes (EUB) estimator of the area mean. For calibration of uncertainty of the EUB estimator, we consider the conditional mean squared error (CMSE) and derive the second-order unbiased estimator motivated from the work of Booth and Hobert (1998), Datta, Kubokawa, Molina and Rao (2011) and Sugasawa and Kubokawa (2016). As typical examples, we handle three models, namely the Fay-Herriot model for continuous data, the Poisson-gamma and binomial-beta models for count data. It is shown that the shrinkage property pointed out in Datta and Mandal (2015) in the Fay-Herriot model still holds in both the Poisson-gamma and binomial-beta models. That is, the shrinkage coefficient in the EUB estimator decreases as the y_i gets close to the synthetic mean.

This paper is organized as follows. In Section 2, we provide the detailed description of the proposed model, the EM algorithm for parameter estimation and three examples. In Section 3, we derive the second order unbiased estimator of CMSE for risk evaluation of the EUB estimator. Simulation studies and empirical applications are given in 4 and 5, respectively.

2 Empirical Uncertain Bayes Methods

2.1 Model setup and uncertain Bayes estimator

Let y_1, \dots, y_m be mutually independent random variables where the conditional distribution of y_i given θ_i belongs to the the following natural exponential family:

$$y_i|\theta_i \sim f(y_i|\theta_i) = \exp\{n_i(\theta_i y_i - \psi(\theta_i)) + c(y_i, n_i)\}, \quad (2)$$

where n_i is a known scalar parameter and is not necessarily corresponding to the sample size in the i th area. As the prior distribution of θ_i , we set the uncertain random structure treated in Datta and Mandal (2015). Let s_1, \dots, s_m be mutually independent and identical random variables distributed as

$$P(s_i = 1) = p, \quad P(s_i = 0) = 1 - p.$$

The prior distribution of θ_i is given by

$$\theta_i|(s_i = 1) \sim \pi(\theta_i) = \exp\{\nu(m_i\theta_i - \psi(\theta_i)) + C(\nu, m_i)\}, \quad \theta_i|(s_i = 0) = (\psi')^{-1}(m_i), \quad (3)$$

where ν is an unknown scalar hyperparameter, $C(\nu, m_i)$ is the normalizing constant and $\psi'(t) = d\psi(t)/dt$. In our settings, we consider the canonical link

$$m_i = \psi'(\mathbf{x}_i^t \boldsymbol{\beta}),$$

where \mathbf{x}_i is a $q \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $q \times 1$ unknown common vector of regression coefficients. The function $f(y_i|\theta_i)$ is the regular one-parameter exponential family and the function $\pi(\theta_i)$ is the conjugate prior distribution. Then the unknown parameter in two-stage model (2) and (3) are $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, \nu, p)^t$. The quantity of interest in this paper is the conditional expectation of y_i given θ_i , defined as

$$\mu_i = E[y_i|\theta_i] = \psi'(\theta_i),$$

noting that $\mu_i(s_i = 0) = m_i$ from (3). For $\psi''(t) = d^2\psi(t)/dt^2$, we assume that $\psi''(\theta_i) = Q(\mu_i)$, namely,

$$\text{Var}(y_i|\theta_i) = \frac{\psi''(\theta_i)}{n_i} = \frac{Q(\mu_i)}{n_i},$$

with $Q(x) = v_0 + v_1x + v_2x^2$ for known constants v_0, v_1 and v_2 which are not simultaneously zero. This means that the conditional variance $\text{Var}(y_i|\theta_i)$ is a quadratic function of the conditional expectation $E[y_i|\theta_i]$. This family is known as a natural exponential family with the quadratic variance function studied by Morris (1982, 1983). Similarly, the mean and variance of the prior distribution given $s_i = 1$ are

$$E[\mu_i|s_i = 1] = m_i, \quad \text{Var}(\mu_i|s_i = 1) = \frac{Q(m_i)}{\nu - v_2}.$$

The joint density (or mass) function of (y_i, θ_i, s_i) is

$$g(y_i, \theta_i, s_i = 1) = f(y_i|\theta_i)\pi(\theta_i), \quad g(y_i, \theta_i, s_i = 0) = \delta_{\theta_i}((\psi')^{-1}(m_i))f(y_i|\theta_i),$$

where $\delta_{\theta_i}(a)$ denotes the point mass at $\theta_i = a$. Then the joint distribution of (y_i, θ_i) and the marginal distribution of y_i are both mixtures of two distributions:

$$\begin{aligned} g(y_i, \theta_i) &= pf(y_i|\theta_i)\pi(\theta_i) + (1-p)\delta_{\theta_i}((\psi')^{-1}(m_i))f(y_i|\theta_i), \\ f(y_i; \boldsymbol{\phi}) &= pf_1(y_i; \boldsymbol{\phi}) + (1-p)f_2(y_i; \boldsymbol{\phi}), \end{aligned}$$

where

$$f_1(y_i; \boldsymbol{\phi}) = \int f(y_i|\theta_i)\pi(\theta_i)d\theta_i, \quad f_2(y_i; \boldsymbol{\phi}) = f(y_i|\theta_i = (\psi')^{-1}(m_i)). \quad (4)$$

Since $\pi(\theta_i)$ is the conjugate prior of θ_i , the marginal distribution $f_1(y_i; \boldsymbol{\phi})$ and the conditional distribution $\pi(\theta_i|y_i, s_i = 1; \boldsymbol{\phi})$ can be obtain in the closed forms:

$$\begin{aligned} \pi(\theta_i|y_i, s_i = 1; \boldsymbol{\phi}) &= \exp \{ (n_i + \nu)(\eta_i\theta_i - \psi(\theta_i)) \} C(n_i + \nu, \eta_i), \\ f_1(y_i; \boldsymbol{\phi}) &= \frac{C(\nu, m_i)}{C(n_i + \nu, \eta_i)} \exp \{ c(y_i, n_i) \}, \end{aligned}$$

where

$$\eta_i \equiv \eta_i(y_i; \boldsymbol{\phi}) = \frac{n_i y_i + \nu m_i}{n_i + \nu}.$$

The conditional distribution of s_i given y_i can be obtained as

$$P(s_i = 1|y_i; \boldsymbol{\phi}) = \frac{pf_1(y_i, \boldsymbol{\phi})}{f(y_i; \boldsymbol{\phi})} = \frac{p}{p + (1-p)f_2(y_i; \boldsymbol{\phi})/f_1(y_i; \boldsymbol{\phi})} = 1 - P(s_i = 0|y_i; \boldsymbol{\phi}).$$

To obtain the Bayes estimator of μ_i , we note that

$$E[\mu_i|s_i, y_i; \boldsymbol{\phi}] = m_i + \frac{n_i}{\nu + n_i}(y_i - m_i)I(s_i = 1), \quad (5)$$

where $I(\cdot)$ is an indicator function. Hence the Bayes estimator of μ_i is

$$\tilde{\mu}_i(y_i, \phi) = E[\mu_i | y_i; \phi] = E[E(\mu_i | s_i, y_i; \phi) | y_i; \phi] = m_i + \frac{n_i}{\nu + n_i}(y_i - m_i)r_i(y_i, \phi), \quad (6)$$

where

$$r_i(y_i, \phi) = P(s_i = 1 | y_i; \phi) = \frac{p}{p + (1 - p)f_2(y_i; \phi)/f_1(y_i; \phi)}. \quad (7)$$

It is observed that $r_i(y_i, \phi)$ increases in p and decreases in the ratio $f_2(y_i; \phi)/f_1(y_i; \phi)$. In what follows, we use the abbreviated notations $\tilde{\mu}_i$ and r_i instead of $\tilde{\mu}_i(y_i, \phi)$ and $r_i(y_i, \phi)$, respectively, when there is no confusion. It is noted that the Bayes estimator (6) can be expressed as

$$\tilde{\mu}_i = y_i - \left\{ 1 - \frac{n_i}{\nu + n_i}r_i(y_i, \phi) \right\} (y_i - m_i),$$

which shrinks the direct estimator y_i toward the regression (or synthetic) part $m_i = \psi'(\mathbf{x}_i^t \boldsymbol{\beta})$, and the shrinkage function depends on y_i through r_i . On the other hand, in the classical two-stage model as used in Ghosh and Maiti (2004), the shrinkage function does not depend on the observation y_i , which is sometimes not flexible for real data analysis. It should be noted that $r_i = 1$ when $p = 1$. Thus, the suggested method includes the classical method as well as it has the shrinkage function adjusted by y_i which arises from introducing the weight parameter p . Moreover, when the prior is completely singular, namely $p = 0$, it follows $r_i = 0$ and the resulting Bayes estimator is m_i . We call the estimator (6) under the prior (3) the uncertain Bayes estimator in order to distinguish from the conventional Bayes estimator.

2.2 Maximum likelihood estimation using EM algorithm

Since the uncertain Bayes estimator (6) depends on the unknown model parameter ϕ , we need to estimate them for practical use. A reasonable method is the maximum likelihood (ML) estimator which maximizes the marginal distribution of y_i . Since the marginal density is the mixture of the two distributions $f(y_i; \phi) = pf_1(y_i; \phi) + (1 - p)f_2(y_i; \phi)$, the ML estimator is the maximizer of the log-likelihood function

$$L(\phi) = \sum_{i=1}^m \log \{pf_1(y_i; \phi) + (1 - p)f_2(y_i; \phi)\}. \quad (8)$$

To compute the ML estimate, we propose Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) which maximizes the objective function (8) iteratively and indirectly. From (2) and (3), the complete log-likelihood function $L^c(\phi)$ given $(y_i, \theta_i, s_i)_{i=1, \dots, m}$ is

$$L^c(\phi) = \sum_{i=1}^m \{n_i(\theta_i y_i - \psi(\theta_i))\} + \sum_{i=1}^m s_i \{\nu(m_i \theta_i - \psi(\theta_i)) + C(\nu, m_i)\} + \sum_{i=1}^m \{s_i \log p + (1 - s_i) \log(1 - p)\}.$$

In the r th iteration, we first compute the expectation of the complete log-likelihood $E^{(r)}[L^c(\phi)]$ at the E-step, where $E^{(r)}$ denotes the expectation with respect to the conditional distributions $(\theta_i, s_i) | y_i$ with hyperparameter values $\phi^{(r)}$. Then the objective function to be maximized at the M-step in the r th iteration is

$$\begin{aligned} Q^{(r)}(\phi) \equiv E^{(r)}[L^c(\phi)] &= \sum_{i=1}^m r_i(y_i, \phi^{(r)}) \left\{ \nu m_i E^{(r)}[\theta_i | s_i = 1] - \nu E^{(r)}[\psi(\theta_i) | s_i = 1] + C(\nu, m_i) \right\} \\ &\quad + \sum_{i=1}^m \left\{ r_i(y_i, \phi^{(r)}) \log p + (1 - r_i(y_i, \phi^{(r)})) \log(1 - p) \right\}, \end{aligned}$$

which yields the updating algorithm as

$$\begin{aligned}
(\boldsymbol{\beta}^{(r+1)}, \nu^{(r+1)}) &= \operatorname{argmax}_{\boldsymbol{\beta}, \nu} \sum_{i=1}^m r_i(y_i, \boldsymbol{\phi}^{(r)}) \left\{ \nu m_i E^{(r)}[\theta_i | s_i = 1] - \nu E^{(r)}[\psi(\theta_i) | s_i = 1] + C(\nu, m_i) \right\} \\
p^{(r+1)} &= \frac{1}{m} \sum_{i=1}^m r_i(y_i, \boldsymbol{\phi}^{(r)}).
\end{aligned} \tag{9}$$

Since the prior distribution of θ_i given $s_i = 1$ is conjugate, the posterior distribution of θ_i given $s_i = 1$ belongs to the same family as the prior distribution, and we can easily generate samples from the distribution in common models as demonstrated in the subsequent section. Hence, the calculation of two expectations $E^{(r)}[\theta_i | s_i = 1]$ and $E^{(r)}[\psi(\theta_i) | s_i = 1]$ given in the E-step is easy to carry out. We summarize the EM algorithm in the following.

Algorithm 1 (EM algorithm). *Iterative,*

1. Set the initial value $\boldsymbol{\phi}^{(0)}$ and $r = 0$.
2. Compute $E^{(r)}[\theta_i | s_i = 1]$ and $E^{(r)}[\psi(\theta_i) | s_i = 1]$ using the current parameter value $\boldsymbol{\phi}^{(r)}$.
3. Update the parameter value as $\boldsymbol{\phi}^{(r+1)}$ based on (9).
4. If the difference between $\boldsymbol{\phi}^{(r)}$ and $\boldsymbol{\phi}^{(r+1)}$ is sufficiently small, then the estimate is given by $\boldsymbol{\phi}^{(r+1)}$. Otherwise, set $r = r + 1$ and go back to Step 2.

Finally, substituting $\hat{\boldsymbol{\phi}}$ into the UB estimator, we get the empirical uncertain Bayes (EUB) estimator

$$\hat{\mu}_i \equiv \tilde{\mu}_i(y_i, \hat{\boldsymbol{\phi}}) = \hat{m}_i + \frac{n_i}{\hat{\nu} + n_i} (y_i - \hat{m}_i) r_i(y_i, \hat{\boldsymbol{\phi}}), \tag{10}$$

where $\hat{m}_i = \psi'(\mathbf{x}_i^t \hat{\boldsymbol{\beta}})$.

2.3 Some examples

Here we provide three typical models often used in practice and investigate properties of the UB estimators with detailed expressions of E-step and M-step in the EM algorithm.

[1] Normal-normal (Fay-Herriot) model. The Fay-Herriot model (Fay and Herriot, 1979) is an area-level model frequently used in small area estimation, given by

$$y_i | \theta_i \sim N(\theta_i, D_i), \quad \theta_i | (s_i = 1) \sim N(\mathbf{x}_i^t \boldsymbol{\beta}, A), \quad i = 1, \dots, m,$$

corresponding to $n_i = D_i^{-1}$, $v_0 = 1$, $v_1 = v_2 = 0$, $\nu = A^{-1}$ and $\psi(\theta_i) = \theta_i^2/2$ in (2) and (3). This model was studied in Datta and Mandal (2015) in terms of Bayesian perspectives. The marginal distributions of $f_1(y_i)$ and $f_2(y_i)$ in (4) are given by

$$f_1(y_i; \boldsymbol{\phi}) = \frac{1}{\sqrt{2\pi(A + D_i)}} \exp\left(-\frac{(y_i - m_i)^2}{2(A + D_i)}\right), \quad f_2(y_i; \boldsymbol{\phi}) = \frac{1}{\sqrt{2\pi D_i}} \exp\left(-\frac{(y_i - m_i)^2}{2D_i}\right), \tag{11}$$

so that $r_i(y_i, p)$ is obtained from (7) as

$$r_i(y_i, p) = p \left\{ p + (1 - p) \sqrt{\frac{A + D_i}{D_i}} \exp\left(-\frac{A(y_i - m_i)^2}{2D_i(A + D_i)}\right) \right\}^{-1},$$

which coincides with the result given in Datta and Mandal (2015). It is clear that $r_i(y_i, p)$ takes small values when y_i is close to m_i , corresponding to the case where y_i is well explained by m_i without random effects.

Regarding the parameter estimation via the EM algorithm in the Fay-Herriot model, the objective function at the M-step is

$$Q^{(r)}(\beta, A) = -\frac{1}{2} \sum_{i=1}^m r_i^{(r)} \left\{ \log A + \frac{1}{A} \left(\theta_i^{(r)} - \mathbf{x}_i^t \beta \right)^2 \right\},$$

where $r_i^{(r)} = r_i(y_i, \beta^{(r)}, A^{(r)}, p^{(r)})$ and $\theta_i^{(r)} = (A^{(r)} y_i + D_i \mathbf{x}_i^t \beta^{(r)}) / (A^{(r)} + D_i)$, so that the updating step for β and A is written as

$$\beta^{(r+1)} = \left(\sum_{i=1}^m r_i^{(r)} \mathbf{x}_i \mathbf{x}_i^t \right)^{-1} \sum_{i=1}^m r_i^{(r)} \theta_i^{(r)} \mathbf{x}_i, \quad A^{(r+1)} = \frac{1}{m} \sum_{i=1}^m \left(\theta_i^{(r)} - \mathbf{x}_i^t \beta^{(r+1)} \right)^2.$$

[2] Poisson-gamma model. Let z_1, \dots, z_m be mutually independent random variables having

$$z_i | \lambda_i \sim \text{Po}(n_i \lambda_i), \quad \lambda_i | (s_i = 1) \sim \text{Ga}(\nu m_i, \nu)$$

where $\lambda_1, \dots, \lambda_m$ are mutually independent, $\text{Po}(\lambda)$ denotes the Poisson distribution with mean λ , and $\text{Ga}(a, b)$ denotes the gamma distribution with density $f(x) \propto x^{a-1} \exp(-bx)$. Let $y_i = z_i / n_i$ and $\log m_i = \mathbf{x}_i^t \beta$ for $i = 1, \dots, m$. Then, the notations in (2) and (3) correspond to $v_1 = 1$, $v_0 = v_2 = 0$ and $\psi(\theta_i) = \exp(\theta_i)$. The marginal distributions of $f_1(y_i)$ and $f_2(y_i)$ are given by

$$f_1(y_i; \phi) = \frac{\Gamma(n_i y_i + \nu m_i)}{\Gamma(n_i y_i + 1) \Gamma(\nu m_i)} \left(\frac{n_i}{n_i + \nu} \right)^{n_i y_i} \left(\frac{\nu}{n_i + \nu} \right)^{\nu m_i}, \quad f_2(y_i; \phi) = \frac{(n_i m_i)^{n_i y_i}}{(n_i y_i)!} \exp(-n_i m_i) \quad (12)$$

where $\Gamma(\cdot)$ denotes a gamma function, so that $r_i(y_i, p)$ is written as

$$r_i(y_i, p) = p \left\{ p + (1-p) \frac{\Gamma(\nu m_i) \exp(-n_i m_i)}{\Gamma(n_i y_i + \nu m_i)} (n_i + \nu)^{n_i y_i + \nu m_i} m_i^{n_i y_i} \nu^{-\nu m_i} \right\}^{-1}. \quad (13)$$

Unlike the Fay-Herriot model, it is not clear when $r_i(y_i, p)$ takes small values as a function of y_i . To see this property, let $h(z_i) = (n_i + \nu)^{z_i + \nu m_i} m_i^{z_i} / \Gamma(z_i + \nu m_i)$. It is noted that $r_i(y_i, p)$ depends on $z_i (= n_i y_i)$ through $h(z_i)$. It follows that

$$\frac{h(z_i + 1)}{h(z_i)} = \frac{n_i m_i + \nu m_i}{z_i + \nu m_i},$$

so that we have $h(z_i) \leq h(z_i + 1)$ for $y_i \leq m_i$ and $h(z_i) \geq h(z_i + 1)$ for $y_i \geq m_i$. Then, when y_i is close to m_i , $h(z_i)$ takes a large value, which results in a small value of $r_i(y_i, p)$. This observation is similar to the case of the Fay-Herriot model.

The objective function at the M-step in the EM algorithm can be expressed as

$$Q^{(r)}(\beta, \nu) = \sum_{i=1}^m r_i(y_i, \beta^{(r)}, \nu^{(r)}, p^{(r)}) \left\{ \nu m_i \log \nu - \log \Gamma(\nu m_i) + \nu m_i \int_0^\infty (\log t) f_\Gamma(t; n_i y_i + \nu^{(r)} m_i^{(r)}, n_i + \nu^{(r)}) dt - \nu \frac{n_i y_i + \nu^{(r)} m_i^{(r)}}{n_i + \nu^{(r)}} \right\},$$

where $f_\Gamma(\cdot; a, b)$ denotes the density function of $\text{Ga}(a, b)$. It should be noted that the integral given in the objective function can be easily calculated by generating samples from $\text{Ga}(n_i y_i + \nu^{(r)} m_i^{(r)}, n_i + \nu^{(r)})$.

[3] Binomial-beta model. Let z_1, \dots, z_m be mutually independent random variables having

$$z_i | p_i \sim \text{Bin}(n_i, p_i), \quad p_i | (s_i = 1) \sim \text{Beta}(\nu m_i, \nu(1 - m_i)),$$

where p_1, \dots, p_m are mutually independent, $\text{Bin}(n, p)$ denotes the binomial distribution and $\text{Beta}(a, b)$ denotes the beta distribution with density $f(x) \propto x^{a-1}(1-x)^{b-1}$. Let $y_i = z_i/n_i$ and $m_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta}) / (1 + \exp(\mathbf{x}_i^t \boldsymbol{\beta}))$ for $i = 1, \dots, m$. Then the notations in (2) and (3) correspond to $v_0 = 0$, $v_1 = 1$ and $v_2 = -1$, $\mu_i = p_i = \exp(\theta_i) / (1 + \exp(\theta_i))$ and $\psi(\theta_i) = \log(1 + \exp(\theta_i))$. The marginal distributions of $f_1(y_i)$ and $f_2(y_i)$ are

$$f_1(y_i; \phi) = \binom{n_i}{n_i y_i} \frac{B(\nu m_i + n_i y_i, n_i(1 - y_i) + \nu(1 - m_i))}{B(\nu m_i, \nu(1 - m_i))}, \quad f_2(y_i; \phi) = \binom{n_i}{n_i y_i} m_i^{n_i y_i} (1 - m_i)^{n_i(1 - y_i)},$$

where $B(\cdot, \cdot)$ denotes a beta function, so that $r_i(y_i, p)$ is written as

$$r_i(y_i, p) = p \left\{ p + (1 - p) \frac{B(\nu m_i, \nu(1 - m_i))}{B(\nu m_i + n_i y_i, n_i(1 - y_i) + \nu(1 - m_i))} m_i^{n_i y_i} (1 - m_i)^{n_i(1 - y_i)} \right\}^{-1}.$$

Using the same arguments as in the Poisson-gamma model, we consider the function $h(z_i) = m_i^{z_i} (1 - m_i)^{n_i - z_i} / B(\nu m_i + z_i, n_i - z_i + \nu(1 - m_i))$. Then the straightforward calculation shows that

$$\frac{h(z_i + 1)}{h(z_i)} = \frac{m_i \{n_i - z_i - 1 + \nu(1 - m_i)\}}{(1 - m_i)(\nu m_i + z_i)},$$

whereby $h(z_i) \leq h(z_i + 1)$ for $y_i \leq m_i(1 - n_i^{-1})$ and $h(z_i) \geq h(z_i + 1)$ for $y_i \geq m_i(1 - n_i^{-1})$. Thus, when y_i is close to m_i , $h(z_i)$ takes a large value, which results in a small value of $r_i(y_i, p)$.

The objective function at the M-step in the EM algorithm is expressed as

$$Q^{(r)}(\boldsymbol{\beta}, \nu) = \sum_{i=1}^m r_i(y_i, \boldsymbol{\beta}^{(r)}, \nu^{(r)}, p^{(r)}) \left\{ \nu m_i \int_0^1 (\log t) f_B(t; a_i^{(r)}, b_i^{(r)}) dt \right. \\ \left. + \nu(1 - m_i) \int_0^1 \log(1 - t) f_B(t; a_i^{(r)}, b_i^{(r)}) dt - \log B(\nu m_i, \nu(1 - m_i)) \right\},$$

where $a_i^{(r)} = n_i y_i + \nu^{(r)} m_i^{(r)}$, $b_i^{(r)} = n_i(1 - y_i) + \nu^{(r)}(1 - m_i^{(r)})$ and $f_B(\cdot; a, b)$ denotes the density function of the beta distribution $\text{Beta}(a, b)$. The two integrals given in the above formula can be easily computed by generating samples from $\text{Beta}(a_i^{(r)}, b_i^{(r)})$.

In Figure 1, we draw the shrinkage function $r_i(y_i, p)$ as a function of y_i for the three models, where $n_i = 10$, $\nu = 10$, $m_i = \psi'(\beta)$ at $\beta = 0$, and the solid, dashed and dotted lines correspond to the three values $p = 0.2, 0.5$ and 0.8 , respectively. It is observed from Figure 1 that the shrinkage function in all the models are actually minimized at $y_i = m_i$ as discussed so far, and converges to 1 as y_i goes away from m_i . Especially, it is interesting to point out that in the Poisson-gamma model, the shrinkage ratio is not symmetric around $y_i = m_i$, while the other two models are symmetric around $y_i = m_i$.

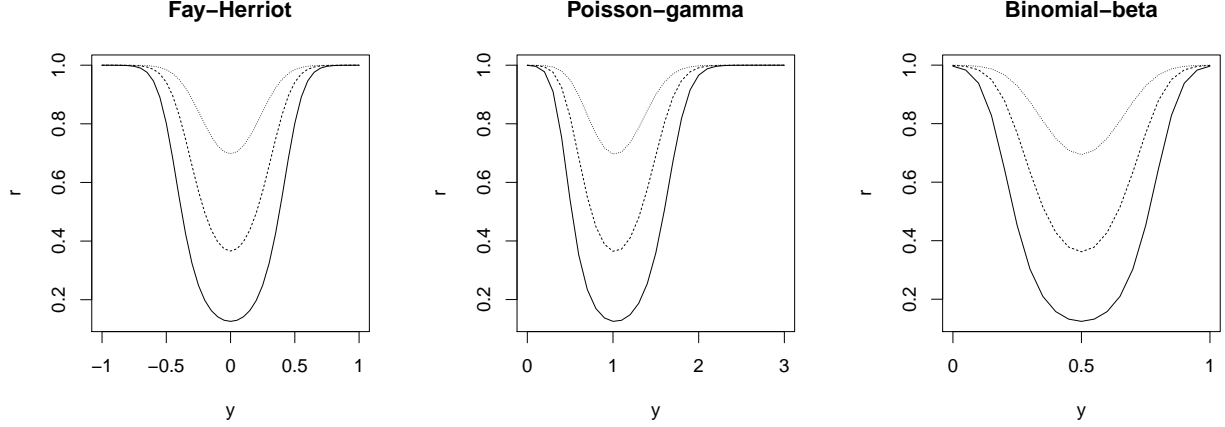


Figure 1: Shrinkage function $r_i(y_i, p)$ in the three models with $p = 0.2$ (solid), 0.5 (dashed), and 0.8 (dotted).

3 Risk Evaluation of the EUB Estimator

3.1 Conditional MSE of the EUB estimator

In practice, the risk evaluation of the resulting estimator is an important issue in small area estimation. The unconditional mean squared error (MSE) is often used, but it is not suitable in this context, because researchers are interested in the risk of the area-specific risk in predicting μ_i under given y_i . This philosophy was originally proposed by Booth and Hobert (1998), and they suggested using the conditional MSE (CMSE) instead of the classical unconditional MSE in the context of mixed model prediction. Since then, the CMSE has been studied in the literature of small area estimation, including Datta, et al. (2011) and Sugawara and Kubokawa (2016). The CMSE of the EUB estimator is defined as

$$\text{CM}_i(y_i, \phi) = \text{E} [(\hat{\mu}_i - \mu_i)^2 | y_i; \phi],$$

noting that the expectation is taken with respect to the conditional distribution $Y_{(-i)} | y_i$ with $Y_{(-i)} = \{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_m\}$. Because $\tilde{\mu}_i$ is the conditional expectation, the CMSE can be decomposed as

$$\text{CM}_i = \text{Var}(\mu_i | y_i; \phi) + \text{E} [(\hat{\mu}_i - \tilde{\mu}_i)^2 | y_i; \phi]. \quad (14)$$

We shall evaluate the two terms in the right hand side of (14).

Concerning the first term in (14), owing to the quadratic variance structure of the assumed model, we have

$$\text{Var}(\mu_i | y_i; \phi) = \text{E} [\text{Var}(\mu_i | s_i, y_i; \phi) | y_i; \phi] + \text{Var} (\text{E}[\mu_i | s_i, y_i; \phi] | y_i; \phi).$$

In the case of $s_i = 1$, we have $\text{Var}(\mu_i | s_i = 1, y_i; \phi) = Q(\eta_i)/(n_i + \nu - v_2)$ for $\eta_i = \text{E}[\mu_i | s_i = 1, y_i; \phi] = (n_i y_i + \nu m_i)/(n_i + \nu)$. Thus,

$$\text{Var}(\mu_i | s_i, y_i; \phi) = \frac{Q(\eta_i)}{n_i + \nu - v_2} I(s_i = 1).$$

From (5), it follows that

$$\text{Var}(\mu_i | y_i; \phi) = \frac{Q(\eta_i)}{n_i + \nu - v_2} P(s_i = 1 | y_i; \phi) + \text{Var} \left(m_i + \frac{n_i}{\nu + n_i} (y_i - m_i) I(s_i = 1) | y_i; \phi \right). \quad (15)$$

Here it is observed that

$$\begin{aligned}\text{Var}\left(m_i + \frac{n_i}{\nu + n_i}(y_i - m_i)I(s_i = 1)|y_i; \phi\right) &= \left(\frac{n_i}{\nu + n_i}\right)^2 (y_i - m_i)^2 E\left[I(s_i = 1) - 2r_i I(s_i = 1) + r_i^2 | y_i; \phi\right] \\ &= \left(\frac{n_i}{\nu + n_i}\right)^2 (y_i - m_i)^2 r_i(1 - r_i),\end{aligned}$$

where r_i is given in (7). We thus get

$$R_{1i}(y_i, \phi) \equiv \text{Var}(\mu_i | y_i; \phi) = \frac{n_i^2}{(\nu + n_i)^2} (y_i - m_i)^2 r_i(1 - r_i) + \frac{r_i Q(\eta_i)}{n_i + \nu - v_2}, \quad (16)$$

which is of order $O_p(1)$.

Concerning the second term $E[(\hat{\mu}_i - \tilde{\mu}_i)^2 | y_i]$ in (14), we approximate it up to second order. For notational simplicity, let $\phi = (\phi_1, \dots, \phi_q, \phi_{q+1}, \phi_{q+2})^t$ for $(\phi_1, \dots, \phi_q)^t = \beta$, $\phi_{q+1} = \nu$ and $\phi_{q+2} = p$. Let $\hat{\phi} = (\hat{\phi}_1, \dots, \hat{\phi}_q, \hat{\phi}_{q+1}, \hat{\phi}_{q+2})^t$ be the ML estimator of ϕ , where $(\hat{\phi}_1, \dots, \hat{\phi}_q)^t = \hat{\beta}$, $\hat{\phi}_{q+1} = \hat{\nu}$ and $\hat{\phi}_{q+2} = \hat{p}$. The asymptotic variance and bias of $\hat{\phi}$ are, respectively, written as

$$\Omega \equiv E\left\{(\hat{\phi} - \phi)(\hat{\phi} - \phi)^t\right\}, \quad B \equiv E(\hat{\phi} - \phi).$$

It is noted that Ω and B are of order $O(m^{-1})$. Assume the following regularity conditions.

Assumption 1.

- (i) *There exist $\underline{n}, \bar{n} > 0$ such that $\underline{n} \leq n_i \leq \bar{n}$ for all $i = 1, \dots, m$.*
- (ii) *The true value of the parameter ϕ is in the interior of Φ , where Φ is the parameter space.*
- (iii) *The densities $f_a(y_i; \phi)$ for $a = 1, 2$ are three times continuously differentiable and satisfies for $j, \ell, k = 1, \dots, q + 2$,*

$$|f_{a(\phi_j)}(y_i; \phi)| + |f_{a(\phi_j \phi_\ell)}(y_i; \phi)| + |f_{a(\phi_j \phi_\ell \phi_k)}(y_i; \phi)| \leq C(y_i, \phi),$$

for fixed ϕ and $E[|C(y_i, \phi)|^{4+\delta}] < \infty$ for some $\delta > 0$.

The assumption (i) is a standard one in this context. For example, in the Fay-Heriot model described in Section 2.3, the assumption corresponds to $\underline{D} \leq D_i \leq \bar{D}$, $i = 1, \dots, m$ for some \underline{D} and \bar{D} , which is usually assumed in the context of small area estimation (e.g. Datta, Rao and Smith, 2005). The assumptions (ii) and (iii) are required for deriving the asymptotic properties of the ML estimator of ϕ as provided in Lemma 1. It should be noted that the typical three models described in Section 2.3 satisfy the assumption (iii), which can be demonstrated in the Appendix.

Since y_1, \dots, y_m are mutually independent, from Theorem 1 in Lohr and Rao (2009), we can get the following lemma about asymptotic properties of estimators.

Lemma 1. *For the ML estimator $\hat{\phi}$, under Assumption 1, it holds that $\sqrt{m}(\hat{\phi} - \phi) = O_p(1)$, $E\left\{(\hat{\phi} - \phi)(\hat{\phi} - \phi)^t | y_i\right\} = \Omega + o_p(m^{-1})$ and*

$$E(\hat{\phi} - \phi | y_i) = B - \Omega L_{i(\phi)}(y_i, \phi) + o_p(m^{-1}),$$

where $L_{i(\phi)}(y_i, \phi) = \partial L_i(y_i, \phi) / \partial \phi$ for $L_i(y_i, \phi) = \log\{pf_1(y_i; \phi) + (1 - p)f_2(y_i; \phi)\}$.

Note that the conditional asymptotic variance of $\hat{\phi}$ does not depend on y_i , while the conditional asymptotic bias depends on y_i . Using Lemma 1, we can evaluate the second term as

$$\begin{aligned} \mathbb{E} [(\hat{\mu}_i - \tilde{\mu}_i)^2 | y_i] &= \mathbb{E} \left[\left\{ \tilde{\mu}_{i(\phi)}^t (\hat{\phi} - \phi) \right\}^2 \middle| y_i \right] + o_p(m^{-1}) \\ &= \text{tr} \left(\Omega \tilde{\mu}_{i(\phi)} \tilde{\mu}_{i(\phi)}^t \right) + o_p(m^{-1}), \end{aligned}$$

where $\tilde{\mu}_{i(\phi)} = \partial \tilde{\mu}_i / \partial \phi$. Let

$$R_{2i}(y_i, \phi) \equiv \text{tr} \left(\Omega \tilde{\mu}_{i(\phi)} \tilde{\mu}_{i(\phi)}^t \right). \quad (17)$$

Since $R_{2i}(y_i, \phi) = O_p(m^{-1})$, we obtain the following theorem.

Theorem 1. Let $\text{CM}_i^*(y_i, \phi) = R_{1i}(y_i, \phi) + R_{2i}(y_i, \phi)$ for R_{1i} and R_{2i} given in (16) and (17), respectively. Under Assumption 1, we have

$$\text{CM}_i(y_i, \phi) = \text{CM}_i^*(y_i, \phi) + o_p(m^{-1}).$$

3.2 Second-order unbiased estimator of CMSE

The approximated CMSE given in Theorem 1 depends on the unknown parameter ϕ , so that it is not feasible in practice. Here we provide a second-order unbiased estimator of the CMSE. In what follows, we use the abbreviated notations R_{1i} and R_{2i} instead of $R_{1i}(y_i, \phi)$ and $R_{2i}(y_i, \phi)$, respectively, without any confusion.

Since $R_{2i} = O_p(m^{-1})$, we estimate it by the plug-in estimator $R_{2i}(y_i, \hat{\phi})$ with second-order unbiasedness, that is, $\mathbb{E}[R_{2i}(y_i, \hat{\phi}) - R_{2i}(y_i, \phi) | y_i] = o_p(m^{-1})$. On the other hand, the plug-in estimator $R_{1i}(y_i, \hat{\phi})$ has a second-order bias, namely $\mathbb{E}[R_{1i}(y_i, \hat{\phi}) - R_{1i}(y_i, \phi) | y_i] = O_p(m^{-1})$, because $R_{1i} = O_p(1)$. To achieve the second-order accuracy, we correct the second-order bias of the estimator $R_{1i}(y_i, \hat{\phi})$. Using the Taylor series expansion, we have

$$R_{1i}(y_i, \hat{\phi}) = R_{1i} + R_{1i(\phi)}^t (\hat{\phi} - \phi) + \frac{1}{2} (\hat{\phi} - \phi)^t R_{1i(\phi\phi)} (\hat{\phi} - \phi) + o_p(m^{-1}),$$

where $R_{1i(\phi)} = \partial R_{1i} / \partial \phi$ and $R_{1i(\phi\phi)} = \partial^2 R_{1i} / \partial \phi \partial \phi^t$. From Lemma 1, it is seen that the second-order bias in $R_{1i}(y_i, \hat{\phi})$ is

$$\begin{aligned} b_i(y_i, \phi) &\equiv R_{1i(\phi)}^t \mathbb{E} \left[(\hat{\phi} - \phi) | y_i \right] + \frac{1}{2} \text{tr} \left(R_{1i(\phi\phi)} \mathbb{E} \left[(\hat{\phi} - \phi)(\hat{\phi} - \phi)^t | y_i \right] \right) \\ &= R_{1i(\phi)}^t (\mathbf{B} - \Omega L_{i(\phi)}) + \frac{1}{2} \text{tr} (R_{1i(\phi\phi)} \Omega). \end{aligned}$$

Thus, the bias-corrected estimator of R_{1i} is given by

$$R_{1i}^{BC}(y_i, \hat{\phi}) = R_{1i}(y_i, \hat{\phi}) - b_i(y_i, \hat{\phi}), \quad (18)$$

which satisfies $\mathbb{E}[R_{1i}^{BC}(y_i, \hat{\phi}) - R_{1i}(y_i, \phi) | y_i] = o_p(m^{-1})$.

Theorem 2. Let $\widehat{\text{CM}}_i = R_{1i}^{BC}(y_i, \hat{\phi}) + R_{2i}(y_i, \hat{\phi})$, where $R_{1i}^{BC}(y_i, \hat{\phi})$ is given in (18). Then, under Assumption 1, we have

$$\mathbb{E} [\widehat{\text{CM}}_i - \text{CM}_i | y_i] = o_p(m^{-1}).$$

To calculate the $\widehat{\text{CM}}_i$, we compute the estimates of $\mathbf{\Omega}$ and \mathbf{B} using the parametric bootstrap method. Let $\widehat{\mathbf{\Omega}}$ and $\widehat{\mathbf{B}}$ be bootstrap estimators of $\mathbf{\Omega}$ and \mathbf{B} , respectively. Then, we have the approximations

$$E[\widehat{\mathbf{\Omega}}] = \mathbf{\Omega} + o(m^{-1}), \quad E[\widehat{\mathbf{B}}] = \mathbf{B} + o(m^{-1}),$$

because $\mathbf{\Omega} = O(m^{-1})$ and $\mathbf{B} = O(m^{-1})$. Moreover, we need to compute $f_{1(\phi)}$, $f_{2(\phi)}$, $R_{1i(\phi)}$, $R_{1i(\phi\phi)}$ in b_i and $\tilde{\mu}_{i(\phi)}$ in R_{2i} at $\phi = \hat{\phi}$. However, their analytical expressions are too complicated to use them in practice. Thus we utilize the numerical derivatives which were suggested in Lahiri, Maiti, Katzoff and Parsons (2007). Let $\{z_m\}$ be a sequence of positive real numbers converging to 0. Based on $\{z_m\}$, we define

$$\begin{aligned} f_{a(\phi_j)}^*(y_i, \hat{\phi}) &= \frac{1}{2z_m} \left\{ f_a(y_i, \hat{\phi} + z_m \mathbf{e}_j) - f_a(y_i, \hat{\phi} - z_m \mathbf{e}_j) \right\}, \quad a = 1, 2 \\ \tilde{\mu}_{i(\phi_j)}^*(y_i, \hat{\phi}) &= \frac{1}{2z_m} \left\{ \tilde{\mu}_i(y_i, \hat{\phi} + z_m \mathbf{e}_j) - \tilde{\mu}_i(y_i, \hat{\phi} - z_m \mathbf{e}_j) \right\} \\ R_{1i(\phi_j)}^*(y_i, \hat{\phi}) &= \frac{1}{2z_m} \left\{ R_{1i}(y_i, \hat{\phi} + z_m \mathbf{e}_j) - R_{1i}(y_i, \hat{\phi} - z_m \mathbf{e}_j) \right\} \end{aligned}$$

where \mathbf{e}_j is a vector of 0's other than the j -th element is 1. Similarly, we define approximations of the second-order partial derivatives of R_{1i} as

$$\begin{aligned} R_{1i(\phi_j \phi_j)}^*(y_i, \hat{\phi}) &= \frac{1}{z_m^2} \left\{ R_{1i}(y_i, \hat{\phi} + z_m \mathbf{e}_j) + R_{1i}(y_i, \hat{\phi} - z_m \mathbf{e}_j) - 2R_{1i}(y_i, \hat{\phi}) \right\}, \quad j = 1, \dots, k \\ R_{1i(\phi_j \phi_\ell)}^*(y_i, \hat{\phi}) &= \frac{1}{2z_m^2} \left[\left\{ R_{1i}(y_i, \hat{\phi} + z_m \mathbf{e}_{j\ell}) + R_{1i}(y_i, \hat{\phi} - z_m \mathbf{e}_{j\ell}) - 2R_{1i}(y_i, \hat{\phi}) \right\} \right. \\ &\quad \left. - z_m^2 \left\{ R_{1i(\phi_j \phi_j)}^*(y_i, \hat{\phi}) + R_{1i(\phi_\ell \phi_\ell)}^*(y_i, \hat{\phi}) \right\} \right], \quad j \neq \ell, \end{aligned}$$

where $\mathbf{e}_{j\ell} = \mathbf{e}_j + \mathbf{e}_\ell$. The justification of the approximations based on these numerical derivatives is given in the following theorem, where the proof is given in the Appendix.

Theorem 3. *Under Assumption 1, we have*

$$\begin{aligned} |f_{a(\phi_j)}^*(y_i, \hat{\phi}) - f_{a(\phi_j)}(y_i, \hat{\phi})| &= O_p(z_m), \quad |\tilde{\mu}_{i(\phi_j)}^*(y_i, \hat{\phi}) - \tilde{\mu}_{i(\phi_j)}(y_i, \hat{\phi})| = O_p(z_m) \\ |R_{1i(\phi_j)}^*(y_i, \hat{\phi}) - R_{1i(\phi_j)}(y_i, \hat{\phi})| &= O_p(z_m), \quad |R_{1i(\phi_j \phi_\ell)}^*(y_i, \hat{\phi}) - R_{1i(\phi_j \phi_\ell)}(y_i, \hat{\phi})| = O_p(z_m) \end{aligned}$$

From Theorem 3, the second-order unbiasedness of the MSE estimator given in Theorem 2 is still valid as far as $z_m = o(m^{-1})$. In our numerical investigation given in the next section, we use $z_m = m^{-5/4}$.

4 Simulation Studies

4.1 Prediction error comparison

We first evaluated a finite sample performance of the proposed empirical uncertain Bayes method. Specifically, we compared the EUB estimator with the traditional empirical Bayes (EB) estimator. We focused on the two models: Poisson-gamma and binomial-beta models described in Section 2.3.

For the Poisson-gamma model, we considered the following data generating process:

$$\text{PG: } (n_i y_i) | \theta_i \sim \text{Po}(n_i \theta_i), \quad \theta_i | (s_i = 1) \sim \text{Ga}(\nu \exp(\beta_0 + \beta_1 x_i), \nu), \quad P(s_i = 1) = p,$$

where $\beta_0 = 0, \beta_1 = 0.5, \nu = 5, m = 50$, n_i 's were generated from the uniform distribution on $\{5, 6, \dots, 30\}$, and x_i 's were generated from a standard normal distribution. The prior probability p takes the values 0.2, 0.4, 0.6, 0.8 and 1, where the conventional Poisson-gamma model corresponds to the data generating process with $p = 1$. We computed both the EUB and EB estimators from the simulated data set. Based on $R = 5,000$ iterations of the data generation, we calculated the mean squared error and the absolute bias which are respectively defined as

$$\text{MSE}_i = \frac{1}{R} \sum_{r=1}^R \left(\hat{\theta}_i^{(r)} - \theta_i^{(r)} \right)^2, \quad \text{Bias}_i = \frac{1}{R} \left| \sum_{r=1}^R \left(\hat{\theta}_i^{(r)} - \theta_i^{(r)} \right) \right|. \quad (19)$$

Define $\text{MSE}_i(\text{EUB})$ and $\text{MSE}_i(\text{EB})$ be the simulated values MSE_i for the EUB estimator and the EB estimator, respectively. Then we computed the ratio $\text{Ra}_i = \text{MSE}_i(\text{EUB})/\text{MSE}_i(\text{EB})$ for each i , and calculated the $q\%$ quantiles of $\{\text{Ra}_1, \dots, \text{Ra}_m\}$ for $q = 5, 25, 50, 75$ and 95. Hence, if Ra_i is smaller than 1, the EUB estimator performs better than the EB estimator in terms of MSE. We similarly define the ratio of the absolute biases, and the results for the five p patterns are given in Table 1. In the scenario $p = 1$, the traditional Poisson-gamma model is the true model and uncertain model is overfitting. However, the results show that the EUB estimator performs as well as the EB estimator, which indicates that the effect of overfitting seems small. Moreover, from Table 1, when p is smaller than 1, it is revealed that the EUB estimator improve the EB estimator in terms of both the MSE and the absolute bias, and the improvement is greater as p gets smaller.

We next compared performances of the two estimators in the binomial-beta model using the data generating process:

$$\text{BB} : (n_i y_i) | \theta_i \sim \text{Bin}(n_i, \theta_i), \quad \theta_i | (s_i = 1) \sim \text{Beta}(\nu m_i, \nu(1 - m_i)), \quad P(s_i = 1) = p,$$

with $m_i = \exp(\beta_0 + \beta_1 x_i) / \{1 + \exp(\beta_0 + \beta_1 x_i)\}$, where $\beta_0 = 0, \beta_1 = 0.5, \nu = 5, m = 50$, n_i 's were generated from the uniform distribution on $\{10, 11, \dots, 30\}$, and x_i 's were generated from a standard normal distribution. Similarly to the previous study, we simulated the MSE and the absolute bias using (19) with $R = 5,000$, and computed the quantiles of the ratios. The results are given in Table 1, which shows the similar results to the Poisson-gamma case. However, the amount of improvement seems smaller than that in the Poisson-gamma case, but the EUB estimator performs better than the EB estimator in the binomial-beta case.

4.2 Sensitivity to distributional assumptions

We next investigated sensitivity to distributional assumptions in the proposed model. Here we focused on the Poisson-gamma model as considered in the previous simulation study:

$$(n_i y_i) | \theta_i \sim \text{Po}(n_i \theta_i), \quad \theta_i | (s_i = 1) \sim \text{Ga}(\nu \exp(\beta_0 + \beta_1 x_i), \nu), \quad P(s_i = 1) = p,$$

where $p = 0.5$, and other settings β, β_1, ν, n_i and x_i are set as the same values as in the previous section. To asses sensitivity of the distributional assumption of the proposed method, we consider the two alternative distributions: a log-normal distribution and a two-point distribution for θ_i instead of the gamma distribution. Noting that $E[\theta_i] = m_i$ and $\text{Var}(\theta_i) = m_i/\nu$ under the gamma distribution, we scaled two distributions to have the same expectation and variance. Specifically, we set $\log \theta_i \sim N(\log(m_i/\sqrt{1+1/\nu m_i}), \log(1+1/\nu m_i))$ for the log-normal distribution, and $P(\theta_i = m_i + \sqrt{m_i/\nu}) = P(\theta_i = m_i - \sqrt{m_i/\nu}) = 0.5$ for the two-point distribution. Based on $R = 5,000$ simulation runs, we computed the MSE and absolute bias with the formula (19) for three underlying distributions. In Table 2, we show the five quantiles of the simulated MSE and absolute bias. It is observed that both

Table 1: Simulated ratios of the MSEs and absolute biases of the EUB estimator over the EB estimator.

		MSE					Absolute bias				
	p	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
PG	0.2	0.156	0.355	0.542	0.717	0.842	0.554	0.628	0.707	0.832	0.878
	0.4	0.298	0.454	0.578	0.710	0.815	0.652	0.693	0.726	0.799	0.874
	0.6	0.540	0.640	0.728	0.826	0.894	0.795	0.821	0.843	0.876	0.911
	0.8	0.700	0.832	0.876	0.912	0.962	0.890	0.911	0.925	0.949	0.977
	1	0.985	0.993	1.000	1.005	1.011	0.989	0.993	0.999	1.002	1.007
BB	0.2	0.395	0.608	0.730	0.807	0.996	0.490	0.583	0.650	0.768	0.980
	0.4	0.488	0.736	0.827	0.887	0.983	0.721	0.747	0.805	0.862	0.987
	0.6	0.730	0.866	0.909	0.934	0.983	0.813	0.851	0.886	0.924	0.994
	0.8	0.865	0.946	0.970	0.984	0.993	0.916	0.938	0.961	0.971	0.992
	1	0.966	0.980	1.001	1.007	1.017	0.979	0.987	0.995	1.004	1.012

the MSE and the absolute bias in the misspecified cases of log-normal and two-point distributions are larger than the correctly specified case of the gamma distribution. The absolute biases in the two misspecified cases are about twice as large as that in the gamma case, so that the inflation of the absolute bias seems relatively large. However, the difference in the MSE is around 10%, so that the effect of misspecification on MSE seems relatively small.

Table 2: Quantiles of the simulated MSE and absolute bias of the EUB estimator under the three underlying distributions (the values are multiplied by 100).

distribution	MSE					Absolute bias				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
Gamma	1.31	2.55	3.98	5.92	11.06	0.03	0.10	0.25	0.48	0.71
Log-normal	1.32	2.47	4.15	5.85	10.89	0.05	0.33	0.49	0.72	1.25
two-point	1.57	2.87	4.50	6.46	10.58	0.08	0.28	0.51	0.84	1.18

4.3 Finite sample performance of the CMSE estimator

Finally we investigated a finite sample behavior of the CMSE estimator provided in Theorem 2 in the UPG model. We considered the simple data generating process without covariates:

$$(n_i y_i) | \lambda_i \sim \text{Po}(n_i \lambda_i), \quad \lambda_i | (s_i = 1) \sim \text{Ga}(\nu \exp(\beta), \nu), \quad P(s_i = 1) = p, \quad (20)$$

with $\beta = 1$, $\nu = 5$, $p = 0.5$ and $n_i = 10$. For the number of areas, we consider the two cases of $m = 50$ and $m = 100$. For conditioning values of y_α , we consider α -quantiles of the marginal distribution of y_i , where $\alpha = 0.1, 0.2, \dots, 0.9$, and calculate these values by generating 10,000 random samples from (20). To get the simulated values of the CMSE given y_α , we generate random samples from (20) and replace y_1 with y_α , and we computed the EUB estimator of λ_1 . For the true values of λ_1 , since y_α is given,

we generate λ_1 from the posterior distribution $\lambda_1|y_\alpha \sim r_1 \text{Ga}(\tilde{\lambda}_1, (n_1 + \nu)^{-1}\tilde{\lambda}_1) + (1 - r_1)\delta_{\lambda_1}(\exp(\beta))$ with $\tilde{\lambda}_1 = (n_1 + \nu)^{-1}(n_1 y_\alpha + \nu \exp(\beta))$ and r_1 given in (13). Then, based on $R = 10,000$ iteration, we calculate the simulated values of the CMSE defined as

$$\text{CM}_\alpha = \frac{1}{R} \sum_{r=1}^R (\hat{\lambda}_1^{(r)} - \lambda_1^{(r)})^2,$$

where $\hat{\lambda}_1^{(r)}$ and $\lambda_1^{(r)}$ are the EUB estimates of λ_1 , and $\lambda_1^{(r)}$ is the generated value from the distribution of $\lambda_1|y_\alpha$ in the r th iteration.

For evaluation of the CMSE estimator, we generated random samples from (20) and replace y_1 with y_α , and get CMSE estimators with $B = 100$ bootstrap samples and $z_m = m^{-5/4}$ for computing the numerical derivatives. This procedure is repeated $S = 2,000$ times and calculated the percentage relative bias (RB) and the coefficient of variation (CV) defined as

$$\text{RB}_\alpha = \frac{1}{S} \sum_{s=1}^S \left(\frac{\widehat{\text{CM}}_\alpha - \text{CM}_\alpha}{\text{CM}_\alpha} \right) \times 100, \quad \text{CV}_\alpha = \sqrt{\frac{1}{S} \sum_{s=1}^S \left(\frac{\widehat{\text{CM}}_\alpha - \text{CM}_\alpha}{\text{CM}_\alpha} \right)^2}.$$

Remember that the suggested CMSE estimator given in Theorem 2 is second-order unbiased. To emphasize the importance of bias correction in estimating the CMSE, we also computed the two criteria of the naive CMSE estimator defined as $\widehat{\text{CM}}_{\alpha(N)} = R_{11}(y_\alpha, \hat{\phi})$. It is noted that the naive estimator has the first order bias, because it ignores the second term R_{2i} and the bias of the plug-in estimator of R_{1i} . For the naive estimator, we calculated RB and CV based on the same number of iteration and we define RBN and CVN as RB and CV of the naive estimator. The resulting values are given in Table 3 for both $m = 50$ and $m = 100$.

It is observed from Table 3, the naive estimator has the serious negative bias when $m = 50$. Especially, when the condition values are upper or lower quantiles, the negative bias tends to be larger. This comes from the fact that the naive estimator ignores the positive $O_p(m^{-1})$ term in the CMSE decomposition given in Theorem 1. Since practitioners decide policies or investments based on estimated values as well as their risk estimates, the under-estimation of the CMSE is considered serious in practice. Hence, the results in Table 3 show that the naive CMSE estimator without bias correction is not suitable for practical use. On the other hand, the bias-corrected CMSE estimator works well in both $m = 50$ and $m = 100$ and provides accurate estimation of the CMSE in terms of the relative biases. Concerning the CV values, the bias-corrected estimator has a slightly larger CV than the naive estimator in most cases. This is because the bias corrected terms increase the variance of the estimator. However, the difference is not so significant. Thus, the bias-corrected CMSE estimator is useful in practice.

5 Illustrative Examples

5.1 Historical mortality data in Tokyo

The mortality rate is a representative index in demographics and has been used in various fields. Especially, in economic history, one can discover new knowledge from a spatial distribution of mortality rate in small areas. As divisions get smaller (e.g. city \rightarrow town \rightarrow block...), one can get a more informative spatial distribution. However, the direct estimate of the mortality rate in small area with extremely low population has high variability, which may leads to incorrect recognition of the spatial distribution. Therefore, it is desirable to use smoothed and stabilized estimates through empirical Bayes methods.

Table 3: Percentage of relative bias and coefficient of variation.

	α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	y_i	1.8	2.1	2.3	2.5	2.7	2.9	3.0	3.3	3.6
$m = 50$	RB	1.21	-1.01	6.87	6.56	5.87	3.43	2.42	-3.95	-9.59
	RBN	-27.7	-23.5	-13.8	-6.56	-3.28	-3.86	-8.37	-21.1	-28.6
	CV	0.53	0.36	0.53	0.67	0.73	0.68	0.64	0.43	0.34
	CVN	0.38	0.37	0.43	0.53	0.61	0.55	0.51	0.39	0.39
$m = 100$	RB	-0.23	-0.45	3.28	3.68	2.02	-0.12	0.55	-2.94	-5.82
	RBN	-16.9	-12.9	-6.33	-3.75	4.79	0.97	-2.55	-12.4	-17.9
	CV	0.28	0.25	0.41	0.57	0.63	0.58	0.52	0.32	0.24
	CVN	0.28	0.27	0.35	0.49	0.52	0.49	0.42	0.29	0.27

We here focus on the mortality data in Tokyo, 1930. The data set consists of the observed mortalities z_i and the number of population N_i in the i th area in Tokyo. Such area-level data are available for $m = 1,371$ small areas. We first computed the expected mortality in the i th area as $n_i = N_i \sum_{j=1}^m z_j / \sum_{j=1}^m N_j$. The standardized mortality ratio (SMR) is defined as the ratio of the actual mortality to the expected mortality for each area, which is often used in epidemiology as an indicator of potential mortality risk (Rothman, Greenland and Lash, 2008). Then, the direct estimator of the SMR in the i th area is $y_i = z_i/n_i$. It is noted that $y_i = 0$ in 84 areas, the number of areas with SMR larger than 1 is 526, and the maximum value of y_i is 16.4.

For this data set, we apply the two models: the uncertain Poisson-gamma model described Section 2.3 and the traditional Poisson-gamma model, described as

$$\begin{aligned} \text{UPG} : n_i y_i | \lambda_i &\sim \text{Po}(n_i \lambda_i), \quad \lambda_i | s_i \sim \text{Ga}(\nu \exp(\beta), \nu), \quad P(s_i = 1) = p \\ \text{PG} : n_i y_i | \lambda_i &\sim \text{Po}(n_i \lambda_i), \quad \lambda_i \sim \text{Ga}(\nu \exp(\beta), \nu), \end{aligned}$$

where $\lambda_i = E(y_i | \lambda_i)$ denotes the ‘true’ SMR in the i -th area, which we want to estimate. Using the EM algorithm in Section 2.2 with 5,000 Monte Carlo samples in each E-step, we get the point estimates of the parameters of the two models as shown in Table 4. For comparison of the two models, we computed AIC and BIC based on the maximum marginal likelihood, and the results are also given in Table 4. In terms of AIC and BIC, the proposed UPG model fits better than the traditional PG model for this data set. This comes from the feature of the data. In the upper left panel of Figure 2, we show the sample plot of the expected mortality n_i and the SMR y_i , noting that the solid line corresponds to the estimated regression line $y_i = \exp(\hat{\beta})$ in the UPG model. It is observed that most y_i are distributed around the regression line, and the random area effects are necessary in most areas. The UPG model tells us about the feature of the data through the estimate of p . The lower left panel of Figure 2 provides a scatter plot of the estimated conditional probability $P(s_i = 1 | y_i)$ and the SMR y_i , where the conditional probability $P(s_i = 1 | y_i)$ corresponds to the probability of existing random area effect in the i th area when y_i is observed. The solid line corresponds to the estimated regression line $y_i = \exp(\hat{\beta})$ in the UPG model. From the figure, we can see that the estimates of $P(s_i = 1 | y_i)$ are dramatically different from area to area, and the probability gets lower as SMR is closer to the regression line. To see the difference of estimated values of λ_i , in the upper right panel of Figure 2, we present the relative differences between estimators from the two models, which are defined as $(\hat{\lambda}_i^{\text{UPG}} - \hat{\lambda}_i^{\text{PG}}) / \hat{\lambda}_i^{\text{PG}}$, where $\hat{\lambda}_i^{\text{UPG}}$ and $\hat{\lambda}_i^{\text{PG}}$ are empirical Bayes estimates of λ_i from the UPG model

and the PG model, respectively. We can observe that the differences are around 10% and are not negligible.

We next calculated the CMSE estimates of the EUB estimates $\hat{\lambda}_i^{\text{UPG}}$ using Theorem 2 with $B = 100$ and $z_m = m^{-5/4}$. For comparison, we also computed the CMSE estimates of $\hat{\lambda}_i^{\text{PG}}$ using Theorem 2 with $p = 1$, $B = 100$ and $z_m = m^{-5/4}$. Then, we computed their difference and their histogram over areas is given in the lower right panel of Figure 2. In the figure, the positive value indicates that the EUB estimator has the smaller CMSE value than the EB estimator, and it is revealed that the EUB estimator can improve the estimation risk over the EB estimator in many areas. In particular, the mean values of CMSEs are 4.2×10^{-2} for UPG and 5.4×10^{-2} for PG, so that the EUB estimator can improve 20% CMSE values over the traditional EB estimator on average.

Finally, we assessed the performance of the two models in terms of prediction accuracy in non-sampled areas. Since areas with small n_i have high variability, we consider to predict y_i of areas with n_i larger than the α -quantile of n_i 's, denoted by q_α . Thus we omitted areas with n_i larger than q_α and computed the estimates of the model parameters using the remaining data. For fixed α , we define the predictive criterion (PC) as

$$\text{PC}_\alpha = \sum_{i=1}^m I(n_i > q_\alpha) (\hat{m}_i - y_i)^2 \bigg/ \sum_{i=1}^m I(n_i > q_\alpha), \quad (21)$$

noting that \hat{m}_i is the best predictor in non-sampled areas. In this example, $\hat{m}_i = \exp(\hat{\beta})$. The values of PC were computed for three quantiles of $\alpha = 0.90, 0.95$ and 0.99 and reported in Table 4. It is revealed that the EUB method can improve PC values over the EB method by about 10%.

Table 4: Point estimates of the model parameters and values of AIC, BIC and PC (multiplied by 100) for the three thresholds.

Estimates	$\hat{\beta}$	$\hat{\nu}$	\hat{p}	AIC	BIC	PC _{0.90}	PC _{0.95}	PC _{0.99}
UPG	-0.039	5.15	0.56	8142.17	8157.84	8.00	7.70	5.69
PG	-0.052	7.42	—	8265.12	8275.57	8.67	8.24	6.15

5.2 Poverty rates in Spanish provinces

We next applied our method to the income data set in Spanish provinces which was used in Molina and Rao (2010), where the data set is available in the R package `sae`. In this application, we focus on estimating area-level poverty rates. We set the poverty level as 0.7 times the median of all the observed incomes, and computed the direct estimates of the poverty rates. As covariates, we calculated area-level rates of female and labors. The scatter plot of the pairs (n_i, y_i) is given in the left panel of Figure 3, from which we can observe that the direct estimate y_i has higher variability as n_i gets smaller.

For the data set, we applied the two models: the uncertain binomial-beta (UBB) model and the traditional binomial-beta (BB) model, described as

$$\begin{aligned} \text{UBB} : n_i y_i | \theta_i &\sim \text{Bin}(n_i, \theta_i), \quad \theta_i | s_i \sim \text{Beta}(\nu m_i, \nu(1 - m_i)), \quad P(s_i = 1) = p \\ \text{BB} : n_i y_i | \theta_i &\sim \text{Bin}(n_i, \theta_i), \quad \theta_i \sim \text{Beta}(\nu m_i, \nu(1 - m_i)), \end{aligned}$$

where y_i is the direct estimate of the true poverty rate p_i , n_i is the number of observations in the i th area, $m_i = \text{logit}(\beta_0 + \beta_1 g_i + \beta_2 f_i)$ for $\text{logit}(x) = \exp(x)/(1 + \exp(x))$, and g_i and f_i are rates of

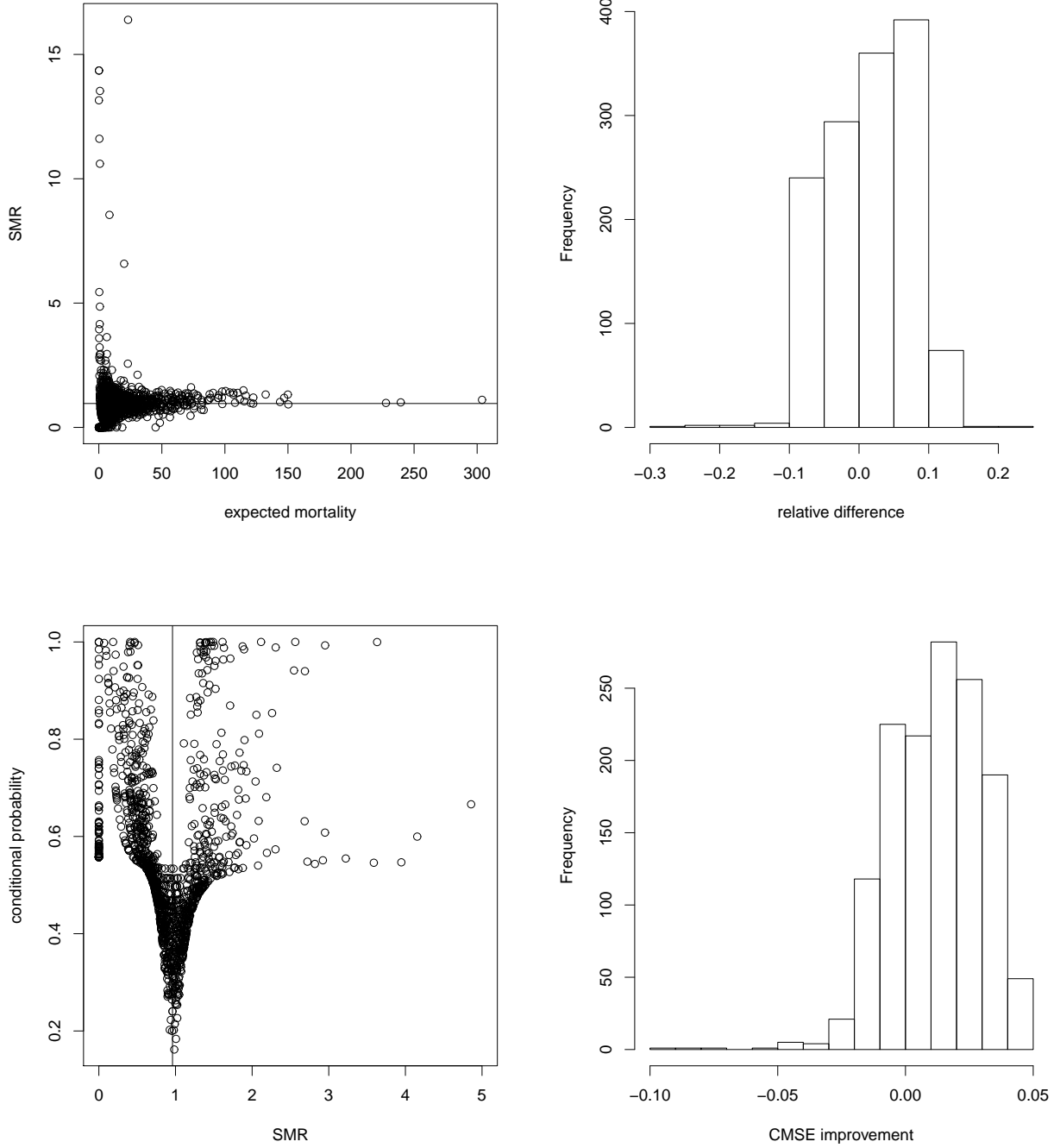


Figure 2: Sample plot of the expected mortality n_i and the SMR y_i (upper-left), the scaled difference of two predictors: $(\hat{\lambda}_i^{\text{UPG}} - \hat{\lambda}_i^{\text{PG}}) / \hat{\lambda}_i^{\text{PG}}$ (upper-right), sample plot of the estimates of conditional probability $P(s_i = 1 | y_i)$ and the SMR y_i (lower-left), and histogram of improvement of estimated CMSE: $\widehat{\text{CM}}_i^{\text{PG}} - \widehat{\text{CM}}_i^{\text{UPG}}$ (lower-right).

populations of females and labors, respectively. The point estimates of the model parameters based on the EM algorithm in Section 2.2 with 5,000 Monte Carlo samples are shown in Table 5. The signs of $\hat{\beta}_1$ and $\hat{\beta}_2$ are reasonable. From the table, it is observed that the estimate of p in the UBB model is almost 1, which implies that the traditional BB model is appropriate for this data set. Actually, the values of AIC and BIC based on the marginal likelihood, given in Table 5, support the BB model rather than the UBB model. Concerning the differences of predicted values, we provide in the right panel of Figure 3 the histogram of the relative differences: $(\hat{\theta}_i^{\text{UBB}} - \hat{\theta}_i^{\text{BB}})/\hat{\theta}_i^{\text{BB}}$, where $\hat{\theta}_i^{\text{UBB}}$ and $\hat{\theta}_i^{\text{BB}}$ are predicted values from the UBB and the BB models. It shows that the differences are smaller than 1% in most areas.

We next calculated the CMSE estimates of the EUB and the EB estimates using Theorem 2 with $B = 100$ and $z_m = m^{-5/4}$. These two estimates are expected to be similar, but the CMSE estimates of the EUB estimates are negative in some areas, while those of the EB estimates are all positive. This may come from the instability of estimating p close to 1.

Finally, we considered the performances of prediction for non-sampled areas. Similarly to the previous section, we considered the predictive criterion (PC) defined in (21) with $\hat{m}_i = \text{logit}(\hat{\beta}_0 + \hat{\beta}_1 g_i + \hat{\beta}_2 f_i)$. The values of PC were computed for $\alpha = 0.70, 0.80$ and 0.90 and reported in Table 5. It is observed that the UBB model provides the performance better than the BB model while the differences are quite small.

Table 5: Point estimates of the model parameters, values of AIC, BIC and PC (multiplied by 1,000) for the three thresholds.

Estimates	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\nu}$	\hat{p}	AIC	BIC	PC _{0.70}	PC _{0.80}	PC _{0.90}
UBB	-1.92	2.91	-1.03	41.33	0.96	459.67	469.42	5.59	5.42	5.17
BB	-2.14	3.36	-1.07	42.93	—	457.74	465.55	5.61	5.42	5.19

6 Concluding Remarks

In this article, we have developed the empirical uncertain Bayes (EUB) methods in area-level models based on the natural exponential family with quadratic variance functions. We have provided the estimators of the model parameters and the EUB estimator as an alternative of the usual empirical Bayes estimator. As examples, we consider the three models often used in practice, namely, the Fay-Herriot, Poisson-gamma and binomial-beta models, and discussed shrinkage properties of the EUB estimator. For the risk evaluation, in the principal of Booth and Hobert (1998), we have constructed the second-order unbiased estimator of the conditional mean squared error. Through simulation studies and empirical applications, we have confirmed that the proposed EUB estimator performs better than the usual empirical Bayes estimator and the suggested CMSE estimator works well in finite samples. Since a hierarchical Bayesian approach enables us to obtain more information about the parameter of interest such as a credible interval, it may be a valuable future study. The data sets and R codes used in Section 5 are available in the supplementary material.

Acknowledgement

We would like to thank the Editor, the Associate Editor and two reviewers for many valuable comments and helpful suggestions which led to an improved version of this paper. The first author

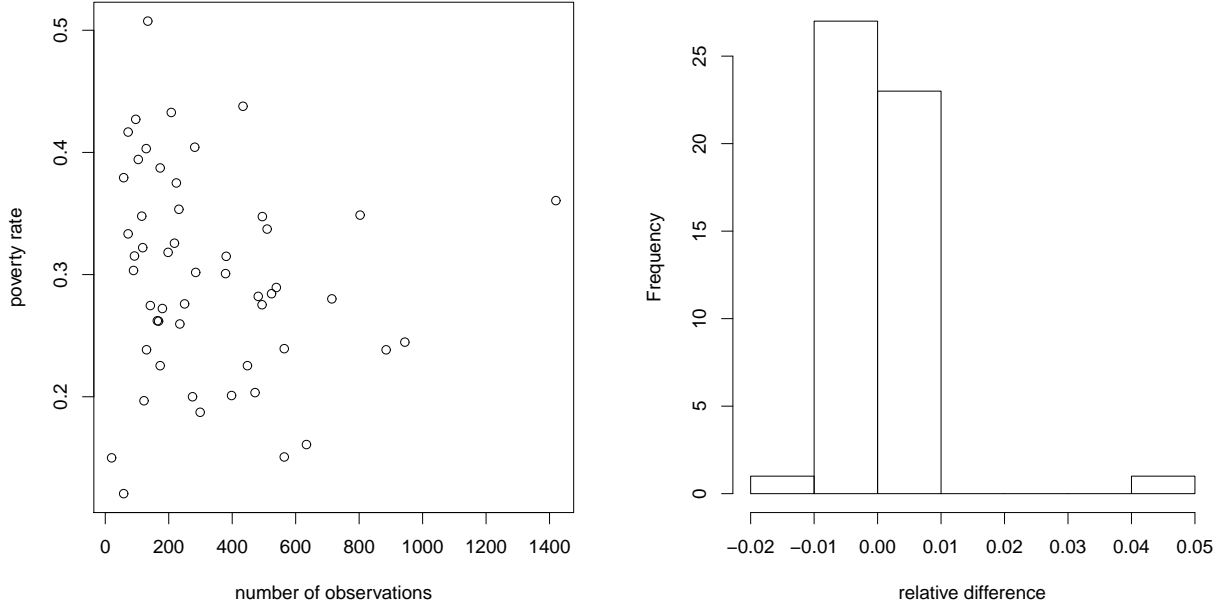


Figure 3: The scatter plot of the number of observations n_i and the direct estimate of poverty rate y_i (left), and the histogram of relative difference $(\hat{\theta}_i^{\text{UBB}} - \hat{\theta}_i^{\text{BB}})/\hat{\theta}_i^{\text{BB}}$ (right).

was supported in part by Grant-in-Aid for Scientific Research (15J10076) from Japan Society for the Promotion of Science (JSPS). Research of the second author was supported in part by Grant-in-Aid for Scientific Research (15H01943 and 26330036) from Japan Society for the Promotion of Science. Research of the third author was supported in part by Grant-in-Aid for Scientific Research (16K17153) from Japan Society for the Promotion of Science.

Appendix

A1. Checking (iii) in Assumption 1.

(Fay-Herriot model). It follows from (11) that

$$f_{1(\beta)}(y_i; \phi) = f_1(y_i; \phi) \left(\frac{y_i - \mathbf{x}_i^t \boldsymbol{\beta}}{A + D_i} \right) \mathbf{x}_i, \quad f_{1(A)}(y_i; \phi) = \frac{f_1(y_i; \phi)}{2(A + D_i)^2} \left\{ (y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2 - A - D_i \right\}.$$

Using $f_1(y_i; \phi) \leq 1/\sqrt{2\pi A}$, we can see that $|f_{1(\phi_j)}(y_i; \phi)|$, $|f_{1(\phi_j \phi_\ell)}(y_i; \phi)|$ and $|f_{1(\phi_j \phi_\ell \phi_k)}(y_i; \phi)|$ can be evaluated from above by 6th order polynomials of y_i and the assumption (iii) is satisfied for $a = 1$. The case of $a = 2$ can be shown similarly.

(Poisson-gamma model). It is noted that $f_{1(\phi_k)}(y_i; \phi) = f_1(y_i; \phi) \partial \log f_1(y_i; \phi) / \partial \phi_k$. From (12),

it holds that

$$\begin{aligned}\frac{\partial \log f_1(y_i; \phi)}{\partial \beta_k} &= \nu x_{ik} m_i \{ \psi(n_i y_i + \nu m_i) - \psi(\nu m_i) \}, \quad k = 1, \dots, q, \\ \frac{\partial \log f_1(y_i; \phi)}{\partial \nu} &= m_i \{ \psi(n_i y_i + \nu m_i) - \psi(\nu m_i) \} - \frac{n_i y_i}{n_i + \nu} + m_i \left\{ \frac{n_i}{n_i + \nu} + \log \left(\frac{n_i}{n_i + \nu} \right) \right\},\end{aligned}$$

where $\psi(\cdot)$ is the digamma function $\psi(x) = d \log \Gamma(x) / dx$. Using the fact that $\psi(x) \approx \log x$ for large x , we have $|\partial \log f_1(y_i; \phi) / \partial \beta_k| = O_p(\log y_i)$ and $|\partial \log f_1(y_i; \phi) / \partial \nu| = O_p(y_i)$ for large y_i . Since there exists $c \in \mathbb{R}$ such that $f_1(y_i; \phi) \leq c$, $|f_1(\phi_k)(y_i; \phi)|$ is bounded above by an liner function of y_i . Concerning the second derivatives, we note that

$$f_{1(\phi_k \phi_\ell)}(y_i; \phi) = f_1(y_i; \phi) \left\{ \frac{\partial^2 \log f_1(y_i; \phi)}{\partial \phi_k \partial \phi_\ell} + \frac{\partial \log f_1(y_i; \phi)}{\partial \phi_k} \frac{\partial \log f_1(y_i; \phi)}{\partial \phi_\ell} \right\}. \quad (22)$$

Moreover, the straightforward calculation shows that

$$\frac{\partial^2 \log f_1(y_i; \phi)}{\partial \beta_k^2} = \nu x_{ik}^2 m_i \{ \psi(n_i y_i + \nu m_i) - \psi(\nu m_i) \} + \nu x_{ik}^2 m_i^2 \left\{ \psi^{(1)}(n_i y_i + \nu m_i) - \psi^{(1)}(\nu m_i) \right\},$$

where $\psi^{(n)}(x) = d^n \psi(x) / dx^n$ is a polygamma function. Since $\psi^{(n)}(x) \approx (-1)^n (n-1)! x^{-n}$ for large x , we have $|\partial^2 \log f_1(y_i; \phi) / \partial \beta_k^2| = O(\log y_i)$ for large y_i . Similarly, we obtain $|\partial^2 \log f_1(y_i; \phi) / \partial \beta_k \partial \nu| = O(\log y_i)$ and $|\partial^2 \log f_1(y_i; \phi) / \partial \nu^2| = O(y_i)$ for large y_i . Thus from expression (22), $|f_{1(\phi_k \phi_\ell)}(y_i; \phi)|$ is bounded above by an quadratic function of y_i . The similar argument shows that $|f_{1(\phi_k \phi_\ell)}(y_i; \phi)|$ is bounded above by an cubic function of y_i . Hence, conditional (iii) is satisfied for $a = 1$, because $E[y_i^c] < \infty$ for all $c > 0$ when y_i has the Poisson-gamma model. The case of $a = 2$ can be shown similarly.

(Binomial-beta model). Note that $f_1(y_i; \phi)$ and $f_2(y_i; \phi)$ have compact supports and the derivatives $f_{a(\phi_j)}(y_i; \phi)$, $f_{a(\phi_j \phi_\ell)}(y_i; \phi)$ and $f_{a(\phi_j \phi_\ell \phi_k)}(y_i; \phi)$ are finite for an interior point ϕ . Then condition (iii) is easy to check.

A2. Proof of Theorem 3. Let us fix ϕ_0 as an interior point of Φ . We here use the notation $C(y_i)$ as a generic function of y_i with $C(y_i) = O_p(1)$, and the notations bu_{1i} and $u_{2i} \in [-1, 1]$ as generic constants. Expanding $f_a(y_i, \phi_0 + z_m \mathbf{e}_j)$ and $f_a(y_i, \phi_0 - z_m \mathbf{e}_j)$ around ϕ_0 , we get

$$\begin{aligned}f_a(y_i, \phi_0 + z_m \mathbf{e}_j) &= f_a(y_i, \phi_0) + f_{a(\phi_j)}(y_i, \phi_0) z_m + \frac{1}{2} f_{a(\phi_j \phi_\ell)}(y_i, \phi_0 + u_{1i} z_m \mathbf{e}_j) z_m^2 \\ f_a(y_i, \phi_0 - z_m \mathbf{e}_j) &= f_a(y_i, \phi_0) - f_{a(\phi_j)}(y_i, \phi_0) z_m + \frac{1}{2} f_{a(\phi_j \phi_\ell)}(y_i, \phi_0 + u_{1i} z_m \mathbf{e}_j) z_m^2,\end{aligned}$$

so that it follows that

$$\begin{aligned}(2z_m) |f_{a(\phi_j)}^*(y_i, \phi_0) - f_{a(\phi_j)}(y_i, \phi_0)| \\ = |f_a(y_i, \phi_0 + z_m \mathbf{e}_j) - f_a(y_i, \phi_0 - z_m \mathbf{e}_j) - 2z_m f_{a(\phi_j)}(y_i, \phi_0)| \\ = \frac{1}{2} z_m^2 |f_{a(\phi_j \phi_\ell)}(y_i, \phi_0 + u_{1i} z_m \mathbf{e}_j) - f_{a(\phi_j \phi_\ell)}(y_i, \phi_0 + u_{2i} z_m \mathbf{e}_j)| \leq C(y_i) z_m^2,\end{aligned}$$

from (iii) of Assumption 1. This shows the first part of Theorem 3.

To show the other parts, we prove that there exist functions $C_a(y_i) = O_p(1)$ for $a = 1, 2, 3$ such that

$$|r_{i(\phi_j)}(y_i, \phi_0)| \leq C_1(y_i), \quad |r_{i(\phi_j \phi_\ell)}(y_i, \phi_0)| \leq C_2(y_i), \quad |r_{i(\phi_j \phi_\ell \phi_k)}(y_i, \phi_0)| \leq C_3(y_i). \quad (23)$$

The straightforward calculation shows that

$$\begin{aligned} r_{i(p)} &= f_1 f_2 \{p f_1 + (1-p) f_2\}^{-2}, \quad r_{i(pp)} = -2 f_1 f_2 \{p f_1 + (1-p) f_2\}^{-3} (f_1 - f_2), \\ r_{i(ppp)} &= 6 f_1 f_2 \{p f_1 + (1-p) f_2\}^{-4} (f_1 - f_2)^2, \end{aligned}$$

which are all bounded above by $C(y_i)$ since $f_a/(p f_1 + (1-p) f_2) \leq \max(p^{-1}, (1-p)^{-1})$ for $a = 1, 2$. Moreover, it is noted that

$$|r_{i(\phi_j)}| = \frac{p(1-p) |f_{1(\psi_j)} f_2 - f_1 f_{2(\psi_j)}|}{\{p f_1 + (1-p) f_2\}^2} \leq \frac{p |f_{1(\psi_j)}| + (1-p) |f_{2(\psi_j)}|}{p f_1 + (1-p) f_2} \leq C(y_i)$$

under (iii) of Assumption 1. Similarly, it can be shown that the higher order derivatives $r_{i(\phi_j \phi_\ell)}$, $r_{i(\phi_j \phi_\ell \phi_k)}$, $r_{i(p \phi_\ell)}$, $r_{i(pp \phi_\ell)}$ and $r_{i(p \phi_j \phi_k)}$ have the form $h(y_i, \phi_0)/\{p f_1 + (1-p) f_2\}^c$, where c is a positive integer and $h(y_i, \phi_0)$ is a polynomial of $f_a, f_{a(\phi_j)}, f_{a(\phi_j \phi_k)}$ and $f_{a(\phi_j \phi_\ell \phi_k)}$, so that there exists $h^\dagger(y_i) = O_p(1)$ such that $h(y_i, \phi_0) \leq h^\dagger(y_i)$. This establishes property (23). Using the property, we have

$$\begin{aligned} (2z_m) |\tilde{\mu}_{i(\phi_j)}^*(y_i, \phi_0) - \tilde{\mu}_{i(\phi_j)}(y_i, \phi_0)| \\ = \frac{1}{2} z_m^2 |\tilde{\mu}_{i(\phi_j \phi_j)}(y_i, \phi_0 + u_{1i} z_m \mathbf{e}_j) - \tilde{\mu}_{i(\phi_j \phi_j)}(y_i, \phi_0 + u_{2i} z_m \mathbf{e}_j)| \leq C(y_i) z_m^2 \end{aligned}$$

and

$$\begin{aligned} (2z_m) |R_{1i(\phi_j)}^*(y_i, \phi_0) - R_{1i(\phi_j)}(y_i, \phi_0)| \\ = \frac{1}{2} z_m^2 |R_{1i(\phi_j \phi_j)}(y_i, \phi_0 + u_{1i} z_m \mathbf{e}_j) - R_{1i(\phi_j \phi_j)}(y_i, \phi_0 + u_{2i} z_m \mathbf{e}_j)| \leq C(y_i) z_m^2. \end{aligned}$$

Finally, we consider the approximation of the second-order partial derivatives of R_{1i} . Expanding $R_{1i}(\phi_0 + z_m \mathbf{e}_j)$ and $R_{1i}(\phi_0 - z_m \mathbf{e}_j)$ up to $O(z_m^3)$, we have

$$\begin{aligned} z_m^2 |R_{1i(\phi_j \phi_j)}^*(y_i, \phi_0) - R_{1i(\phi_j \phi_j)}(y_i, \phi_0)| \\ = \frac{1}{6} z_m^3 |R_{1i(\phi_j \phi_j \phi_j)}(y_i, \phi_0 + u_{1i} z_m \mathbf{e}_j) - R_{1i(\phi_j \phi_j \phi_j)}(y_i, \phi_0 + u_{2i} z_m \mathbf{e}_j)| \leq C(y_i) z_m^3, \end{aligned}$$

from property (23). From this result, we obtain for $j \neq \ell$,

$$\begin{aligned} R_{1i(\phi_j \phi_\ell)}^*(y_i, \phi_0) &= \frac{1}{2z_m^2} \left[\{R_{1i}(y_i, \phi_0 + z_m \mathbf{e}_{j\ell}) + R_{1i}(y_i, \phi_0 - z_m \mathbf{e}_{j\ell}) - 2R_{1i}(y_i, \phi_0)\} \right. \\ &\quad \left. - z_m^2 \{R_{1i(\phi_j \phi_j)}(y_i, \phi_0) + R_{1i(\phi_\ell \phi_\ell)}(y_i, \phi_0)\} \right] + O_p(z_m). \end{aligned}$$

It is noted that

$$\begin{aligned} R_{1i}(y_i, \phi_0 + z_m \mathbf{e}_{j\ell}) &= R_{1i}(y_i, \phi_0) + R_{1i(\psi_j)}(y_i, \phi_0) z_m + R_{1i(\phi_\ell)}(y_i, \phi_0) z_m + R_{1i(\phi_j \phi_\ell)}(y_i, \phi_0) z_m^2 \\ &\quad + \frac{1}{2} R_{1i(\phi_j \phi_j)}(y_i, \phi_0) z_m^2 + \frac{1}{2} R_{1i(\phi_\ell \phi_\ell)}(y_i, \phi_0) z_m^2 + \frac{1}{6} z_m^3 \sum_{j, \ell, k} R_{1i(\phi_j \phi_\ell \phi_k)}(y_i, \phi_0 + z_m u_{1j\ell k} \mathbf{e}_{j\ell k}) \end{aligned}$$

where $u_{1j\ell k} \in [-1, 1]$ and $\mathbf{e}_{j\ell k} = \mathbf{e}_j + \mathbf{e}_\ell + \mathbf{e}_k$. Then it follows that

$$\begin{aligned} z_m^2 |R_{1i(\phi_j \phi_\ell)}^*(y_i, \phi_0) - R_{1i(\phi_j \phi_\ell)}(y_i, \phi_0)| \\ = \frac{1}{6} z_m^3 \left| \sum_{j, \ell, k} R_{1i(\phi_j \phi_\ell \phi_k)}(y_i, \phi_0 + z_m u_{1j\ell k} \mathbf{e}_{j\ell k}) - \sum_{j, \ell, k} R_{1i(\phi_j \phi_\ell \phi_k)}(y_i, \phi_0 + z_m u_{2j\ell k} \mathbf{e}_{j\ell k}) \right|, \end{aligned}$$

for some $u_{2j\ell k} \in [-1, 1]$. Using property of (23), we conclude that the above term is bounded above by $C(y_i) z_m^3$, which completes the proof.

References

- [1] Booth, J. S. and Hobert, P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, **93**, 262-272.
- [2] Datta, G.S., Kubokawa, T., Molina, I. and Rao, J.N.K. (2011). Estimation of mean squared error of model-based small area estimators, *TEST*, **20**, 367-388.
- [3] Datta, G. S. and Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, **110**, 1735-1744.
- [4] Datta, G.S., Rao, J.N.K. and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92**, 183-196.
- [5] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society: Series B*, **39**, 1-38.
- [6] Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.
- [7] Ghosh, M. and Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*, **91**, 95-112.
- [8] Lahiri, A. N., Maiti, T., Katzoff, M and Parsons, V. (2007). Resampling-based empirical prediction: an application to small area estimation. *Biometrika*, **94**, 469-485.
- [9] Lohr, S. L. and Rao, J. N. K. (2009). Jackknife estimation of mean squared error of small are predictors in nonlinear mixed models. *Biometrika*, **96**, 457-468.
- [10] Molina, I. and Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, **38**, 369-385.
- [11] Morris, C. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, **10**, 65-80.
- [12] Morris, C. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics*, **11**, 515-529.
- [13] Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation, 2nd Edition*. Wiley.
- [14] Rothman, K.J., Greenland, S. and Lash, T.L. (2008). *Modern Epidemiology, 3rd Edition*. Lippincott Williams and Wilkins.
- [15] Sugawara, S. and Kubokawa, T. (2016). On conditional prediction errors in mixed models with application to small area estimation. *Journal of Multivariate Analysis*, **148**, 18-33.